SOVEREIGN: MMBERT: Scaled Mixture-of-Experts Multimodal BERT for Robust Chinese Hate Speech

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Hate speech detection on Chinese social networks presents distinct challenges, particularly due to the widespread use of cloaking techniques designed to evade conventional text-based detection systems. Although large language models (LLMs) have recently improved hate speech detection capabilities, the majority of existing work has concentrated on English datasets, with limited attention given to multimodal strategies in the Chinese context. In this study, we propose MMBERT, a novel BERT-based multimodal framework that integrates textual, speech, and visual modalities through a Mixture-of-Exper

1 Introduction

Analysis of: MMBERT: Scaled Mixture-of-Experts Multimodal BERT for Robust Chinese Hate Speech Detection under Cloaking Perturbations. Research goal: Does SMoES's routing strategy enhance robustness to cross-modal distribution shifts on the MMMU benchmark (e.g., visual vs. textual adversarial perturbations) relative to dense models and modality-agnostic MoE baselines across different model scales?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

1 papers retrieved. 3 claims extracted, 3 verified. Tribunal: 7.5/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
MMBERT is a novel BERT-based multimodal framework that integrates textual, speech, and visual modalities through a Mixtu	✓	0.36
MMBERT incorporates modality-specific experts, a shared self-attention mechanism, and a router-based expert allocation s	✓	0.34
Empirical results in several Chinese hate speech datasets show that MMBERT significantly surpasses fine-tuned BERT-based	✓	0.46

References

- https://www.semanticscholar.org/paper/89848074de9ff39536cbbb1f7c3d810f15b6f81a